Homework 4

Midway Check-in Final Project

Due: July 30, Sunday, 11:59 pm

Instructions

- This Homework 4 serves as a midpoint evaluation for your Final Project.
- Combined with the Final Project, it will assess your overall understanding of the course, the fulfillment of course objectives, and the achievement of learning outcomes.
- You will be working on an individual data science project from start to finish.
- Remember, both the course instructor and TA Bennett are available to assist you with any challenges you encounter. Office hours are available by appointment too. If in doubt, ask.
- The project does not have a predefined "solution"; you will be evaluated based on your approach and implementation of the data science project.
- Please refer to the shared grading criteria and project structure below for evaluation details.

Grade Distribution

Submission	Total	% Contribution to Final Grade	Impact of 1 point of Submission in Final Grade
HW1	14	10.5	0.75
HW2	50	19.5	0.39
HW3	50	20	0.40
HW4	50	20	0.40
Final Project	-	20	-
Weekly Checkin	14	5	0.35
Participation Activities	5	5	0.20
Total	-	100	-
Bonus Questions	5	5	1

Table 1: Grade Distribution

1 Setup

1.1 Dataset Selection

Two options:

1. Pick from an existing list
   (a) Access the Google sheet URL on D2L, where the datasets are shared.
   (b) Select any dataset that hasn't been chosen by another student and write your name on the next cell to claim it. Ensure that the dataset can solve some form of problem. Add what problem the dataset solves into the third column.
2. Find a dataset
   (a) Find a dataset that do not have a lot of work/solutions around the dataset.
   (b) Email the link to Bennett with what problem the dataset solves.cc Enfa.
   (c) Once approved, add it to the sheets shared in D2L.

1.2 Github Setup

1. Visit the GitHub project URL that is shared on the website.
2. You will be asked to accept the homework
3. Instruction on the screen will guide you to clone a repo.
4. Clone the repo to your local ie computer.
5. Commit and push your code whenever you complete some form of a unit of the homework.

   For example: After I complete data visualizations, I may add the changes with a commit message "add: Data Viz" and push it. This way, you won't lose your work from unexpected accidents, and the instructor and TA can help you debug easier since we have access to it too.
6. Do NOT make the repo public until the results of finals are released.

2 Data Collection [5 points]

2.1 Dataset Description [5 points]

Write a short description of the dataset (200-word minimum). Below are some questions that can serve as a starting point. You may not have answers to all of these. These questions are meant to serve only as a guide.

- Where was the data collected from?
- How did they collect the data?
- Whose data is this?
- What are the main columns/features of the dataset? What does each of the colum represent? You can add a table here if you want to.
- How old is it?
- Does this dataset get updated often? What is the update frequency?

2.2 Dataset Download [Incomplete Penalty -2 points]

- Download the dataset to the folder data.
- Add data to the git ignore file. This way, git will not add your dataset to your pushed repository. Since it is not our data to upload and distribute, we should not upload it to our repo.

3 Data Preprocessing [20 points]

3.1 Data cleaning [5 points]

1. Perform some initial data exploration to understand the dataset better.
2. Address missing data, if any, by appropriately identifying and quantifying missing values in the dataset.
3. Apply suitable techniques to handle missing values, such as imputation or exclusion, based on a logical approach.
4. Justify the chosen method for dealing with missing data in at least 50 words.
5. If you can find no missing values, this section will still contain the code where you are checking for missing values.

3.2 Outlier Detection [5 points]

1. Check for outliers. You may need data visualization here. Go for a method you see fit.
2. If you do find outliers, apply effective outlier treatment techniques, such as removal, transformation, or capping, depending on the data and context.
3. Explained the rationale behind the chosen outlier handling approach.
4. If you can find no outliers, this section will still contain the code where you are checking for outliers.

3.3 Data Quality and Consistency [5 points]

1. Ensured data quality by verifying data consistency and accuracy.
2. Check for any duplicate records or potential errors and correct them as needed.
3. If you did perform some transformation on the dataset, in at least 25 words, explain what and why?

3.4 Data Preparation for Model training [5 points]

1. Convert the dataset into a form that a model can ingest.
2. This may include shuffling the dataset and splitting it into train, test, and validation.
3. Label encoding may be needed.
4. You may also have to normalize data.
5. Whatever transformation you performed to the data, explain why you did so in at least 25 words.

4 Data Visualisation [8 points]

Choose any two charts/plots etc to visualize the dataset and its properties. For each viz,

- Ensure your plot is clear and has all necessary labels marked.
- Write a short desc of why you choose the viz and what you understood about the dataset from it.

5 Model Training [15 points]

5.1 Picking a model [5 points]

1. Pick any machine learning algorithm that you think works for your problem and data. It needn't necessarily be an algorithm taught in class.

2. Write a short paragraph explaining how the algorithm/model works. I am not looking for a full stats-oriented explanation (even though you are welcome to do so). Your explanation should help your classmates understand how the model basically works.

5.2 Model Training [10 points]

1. Train your model and report your model performance on your validation dataset.

2. Choose appropriate tasks that go along with your model. Examples below.

    - If you are using a neural network model, show your training and validation loss plotted as a graph.
    - If you are using K-means clustering, plot the elbow curve. Choose an appropriate K.

3. List the hyperparameters of the model.

5.3 Report Results [3 points]

Report your results on the test dataset. Write a conclusion to your project.

6 Notebook Presentation [2 points]

While a clean with all outputs displayed notebook gets a credit of 2 points, unorganized and difficult-to-read notebooks get a penalty of 5 points.